PATENT ABSTRACTS OF JAPAN

(11) Publication number : 07-319871
(43) Date of publication of application : 08.12.1995

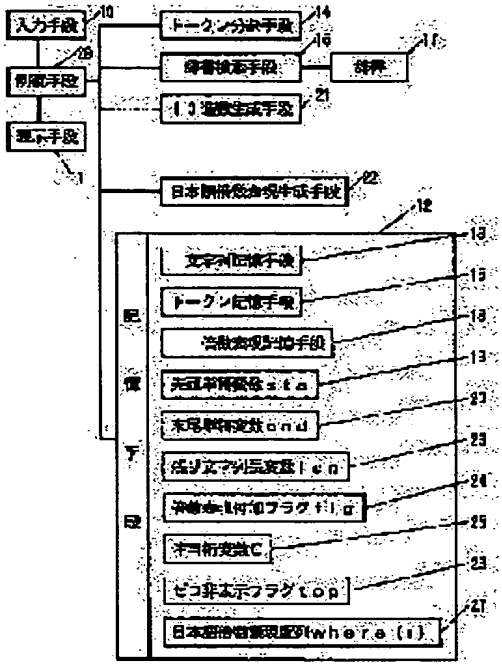
(51) Int.Cl. G06F 17/27

(21) Application number : 06-108362
(22) Date of filing : 23.05.1994

(71) Applicant : MATSUSHITA ELECTRIC IND CO LTD
(72) Inventor : KINOSHITA HITOMI

(54) MORPHEME ANALYSIS DEVICE

(57) Abstract:
PURPOSE: To precisely analyze the expression of an English multiple.
CONSTITUTION: An input means 10 inputting an English character string, character string storage means 13 storing the English character string inputted from the input means 10, a token division means 14 dividing the English character string stored in the character string storage means 13 for respective tokens and a token storage means 15 storing the tokens divided by the token division means 14 are provided. A multiple expression storage means 18 storing the expression of the English multiple and a decimal number generation means 21 generating a decimal number from the token expressing a numerical value by referring to the tokens stored in the token storage means 15 and the expression of the English multiple stored in the multiple expression storage means 18 are provided.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平7－319871

(43)公開日　平成７年(1995)12月８日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｆ　17/27				
		8219－5Ｌ	Ｇ０６Ｆ　15/ 38	Ｅ

審査請求　未請求　請求項の数２　ＯＬ　（全７頁）

(21)出願番号　特願平6－108362

(22)出願日　平成６年(1994)５月23日

(71)出願人　000005821
松下電器産業株式会社
大阪府門真市大字門真1006番地

(72)発明者　木下　ひとみ
大阪府門真市大字門真1006番地　松下電器産業株式会社内

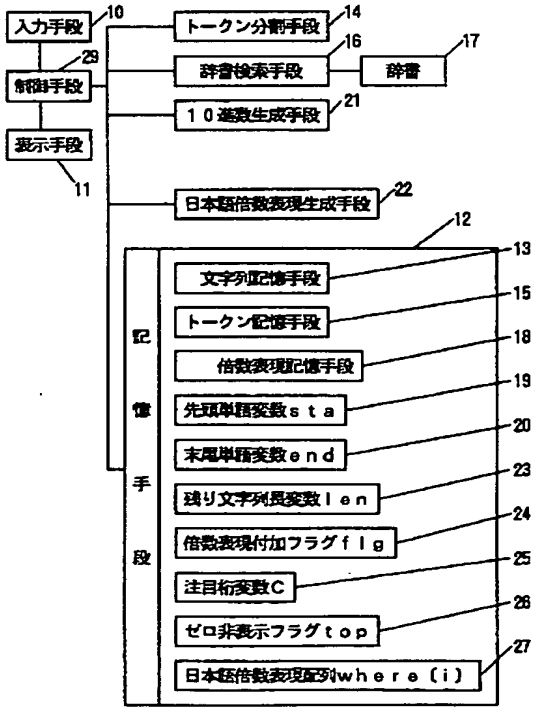
(74)代理人　弁理士　小鍛治　明　（外２名）

(54)【発明の名称】　形態素解析装置

(57)【要約】

【目的】　英語の倍数表現を正しく解析する。

【構成】　英文字列を入力する入力手段１０と、入力手段１０から入力された英文字列を記憶する文字列記憶手段１３と、文字列記憶手段１３に記憶された英文字列をトークン毎に分割するトークン分割手段１４と、トークン分割手段１４により分割されたトークンを記憶するトークン記憶手段１５とを備え、英語倍数表現を記憶する倍数表現記憶手段１８と、トークン記憶手段１５に記憶されたトークンと倍数表現記憶手段１８に記憶された英語倍数表現とを参照して、数値を表現するトークンから１０進数を生成する１０進数生成手段２１とを有する。



1

【特許請求の範囲】

【請求項１】原文文字列を入力する入力手段と、前記入力手段から入力された原文文字列を記憶する文字列記憶手段と、前記文字列記憶手段に記憶された原文字列をトークン毎に分割するトークン分割手段と、前記トークン分割手段により分割されたトークンを記憶するトークン記憶手段とを備え、

原語倍数表現を記憶する倍数表現記憶手段と、前記トークン記憶手段に記憶されたトークンと前記倍数表現記憶手段に記憶された原語倍数表現とを参照して、数値を表現するトークンから、そのトークンに相当する１０進数を生成する１０進数生成手段とを有することを特徴とする形態素解析装置。

【請求項２】前記１０進数生成手段が生成した１０進数に日本語倍数表現を付加する日本語倍数表現生成手段を有することを特徴とする請求項１記載の形態素解析装置。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は入力した原文文字列の形態情報を出力する形態素解析装置に関するものである。

【０００２】

【従来の技術】形態素解析機能を備えた機械翻訳装置により、原文文字列を翻訳するに先立ち、入力した原文文字列をトークン毎に分割し、各トークンに対応する辞書情報を獲得する形態素解析が行われる。

【０００３】

【発明が解決しようとする課題】さて従来の形態素解析装置では、原文を英文として例をあげると、次に述べるように、入力した英文字列に英語倍数表現が含まれていると誤まった形態素情報が出力され、誤訳の原因となるという問題点があった。例えば「ｔｗｏ　ｈｕｎｄｒｅｄ　ａｎｄ　ｔｗｅｎｔｙ－ｔｗｏ」という英語倍数表現が英文字列に含まれているとき、従来の形態素解析装置ではスペースを区切り文字として区切り文字に挟まれた英文字をトークンとするので、この英文字列は、「ｔｗｏ」、「ｈｕｎｄｒｅｄ」、「ａｎｄ」、「ｔｗｅｎｔｙ－ｔｗｏ」の４つのトークンに分割される。そして、このまま機械翻訳装置により翻訳を行うと、「にひゃくと　にじゅうに」ないし「にひゃく　と　にじゅうに」という訳語が得られる。もちろん、この際、正しい訳は「にひゃくにじゅうに（二百二十二、２２２）」である。ここで、英語では「ｈｕｎｄｒｅｄ」、「ｔｈｏｕｓａｎｄ」、「ｍｉｌｌｉｏｎ」、・・・というような倍数表現の次に「ａｎｄ」が付されることがあり、従来の形態素解析装置ではこのような「ａｎｄ」が英文字列に含まれていると、この「ａｎｄ」の前後を分割してしまうものであった。また以上、基数詞について述べたが、「ｆｉｒｓｔ」、「ｓｅｃｏｎｄ」、・・・というような序数詞についても同様の問題点がある。

2

【０００４】そこで本発明は、倍数表現を含む原文文字列を正しく形態素解析することができる形態素解析装置を提供することを目的とする。

【０００５】

【課題を解決するための手段】本発明の形態素解析装置は、原文文字列を入力する入力手段と、入力手段から入力された原文文字列を記憶する文字列記憶手段と、文字列記憶手段に記憶された原文字列をトークン毎に分割するトークン分割手段と、トークン分割手段により分割されたトークンを記憶するトークン記憶手段とを備え、原語倍数表現を記憶する倍数表現記憶手段と、トークン記憶手段に記憶されたトークンと倍数表現記憶手段に記憶された原語倍数表現とを参照して、数値を表現するトークンから、そのトークンに相当する１０進数を生成する１０進数生成手段とを有する。

【０００６】

【作用】上記構成により、入力された原文文字列に倍数表現が含まれている際、この倍数表現が、倍数表現記憶手段及び１０進数生成手段により、１０進数に変換される。

【０００７】

【実施例】次に図面を参照しながら本発明の実施例を説明する。図１は本発明の一実施例における形態素解析装置のブロック図、図２は本発明の一実施例における形態素解析装置の機能ブロック図である。

【０００８】尚、本実施例においては、英文を原文として説明する。図１において、１は図４～図８のフローチャートに沿う制御プログラムを記憶するＲＯＭ（リードオンリーメモリ）、２はＲＯＭ１の制御プログラムを実行し他の各要素を制御するＣＰＵ（中央処理装置）、３は後述する各記憶領域などが設けられているＲＡＭ（ランダムアクセスメモリ）、４は原稿を読取り文字コードに変換して出力するＯＣＲ（光学式文字読取装置）、５はユーザが必要な情報を入力するためのキーボード、６はユーザに必要な情報を表示するためのＣＲＴ（カソードレイチューブ）、７は辞書が設けられたハードディスク装置である。

【０００９】図２において、１０は英文字列を入力する入力手段であり、図１のＯＣＲ４、キーボード５がこれに対応する。１１は処理状況などを表示する表示手段であり、図１のＣＲＴ６がこれに対応する。１２はＲＡＭ３に設けられる記憶手段であり、１３は入力手段１０から入力された英文字列を記憶する文字列記憶手段、１４は文字列記憶手段１３内の英文字列を、区切り文字（スペース、カンマ、セミコロン、コロン、感嘆符、疑問符、カッコなど）を参照しながらトークンに分割するトークン分割手段、１５は分割されたトークン群を記憶するトークン記憶手段、１６はハードディスク装置７に格納された辞書１７を検索する辞書検索手段、１８は図３に示すような英語倍数表現パターン、倍数（詳細は後

述）などを記憶する倍数表現記憶手段、１９は数値を表現するトークンの先頭単語が代入される先頭単語変数ｓｔａ、２０は同末尾単語が代入される末尾単語変数ｅｎｄ、２１はトークン記憶手段１５に記憶されたトークンと倍数表現記憶手段１８に記憶された原語倍数表現とを参照して数値を表示するトークンから１０進数を生成する１０進数生成手段、２２は１０進数生成手段２１が生成した１０進数に「億」、「万」等の日本語倍数表現を付加する日本語倍数表現生成手段である。また２３は日本語倍数表現生成手段２２が駆動される際、１０進数のうち未だ処理をされていない残り文字列の長さを示す残り文字列変数ｌｅｎ、２４は日本語の倍数表現を付加すべきとき「１」、そうでないとき「０」の値を持つ倍数表現付加フラグｆｌｇ、２５は１０進数のうち現在注目している注目桁を示す注目桁変数Ｃ、２６は「億」などの日本語の倍数表現に後続する０を表示すべきでないとき「１」、表示すべきとき「０」の値を持つゼロ非表示フラグｔｏｐである。ここで、日本語の倍数表現に続く０は通常表示しない（例えば、「１億００３０万」ではなく「１億３０万」と表示する）という慣例があるが、このゼロ非表示フラグｔｏｐはこの慣例に従った表示を実現するためのものである。また２７は１０進数を日本語の倍数表現を用いて表示した文字列としての日本語倍数表現配列ｗｈｅｒｅ〔ｉ〕である。ここで、カウンタｉは日本語倍数表現を用いた文字列の先頭文字からの文字数の値である。

【００１０】図３（ａ）は、本発明の一実施例における英語倍数表現パターンの構成図、図３（ｂ）は本発明の一実施例における倍数表現英単語とその倍数との関係図であり、いずれも倍数表現記憶手段１８に格納されている。図３（ａ）中、ＣＡＲＤは基数詞（例えば、「ｔｗｏ」、「ｔｗｅｎｔｙ－ｔｗｏ」など、但し、倍数表現英単語を除く）、Ａ$は「ｈｕｎｄｒｅｄ」、「ｔｈｏｕｓａｎｄ」、・・・などの倍数Ｂ（１０のべき乗）を示す倍数表現英単語、ＯＲＤは「ｆｉｒｓｔ」、「ｓｅｃｏｎｄ」、「ｔｈｉｒｄ」、・・・などの序数詞である。そして倍数表現英単語を用いた英語の数値表現は、図３（ａ）の各パターンのいずれかに該当する。

【００１１】次に図４～図９を参照しながら、本実施例における形態素解析装置の処理の流れについて説明する。ここで図４～図８は、本発明の一実施例における形態素解析装置のフローチャート、図９（ａ）～（ｃ）は本発明の一実施例における形態素解析装置の処理過程説明図である。まず図４は処理の概要を示している。

【００１２】まずステップ１にて、英文字列が入力され文字列記憶手段１３に格納される。ここでは図９（ａ）に示すように「Ｉ　ｓａｗ　ｔｗｅｎｔｙ－ｔｗｏ　ｈｕｎｄｒｅｄ　ａｎｄ　ｆｉｖｅ　ｐｅｏｐｌｅｓ　ｉｎ　ｔｈｅ　ｈｏｌｌ．」という英文字列が入力されたものとする。次に、トークン分割手段１４がこの英文字

列をトークン毎に分割したトークンリストを作成しトークン記憶手段１５に格納する（ステップ２）。ここでは、トークンリストは図９（ｂ）に示すように、トークン毎に前後双方向のポインタ（矢印）で連結した構成となっており、先頭のトークン「Ｉ」と末尾のトークン「．」の一方のポインタは便宜上「ＮＵＬＬ」を指すこととしている。

【００１３】次にステップ３にて各トークンについて辞書情報が獲得され、ステップ４にて図５に示す倍数表現処理（詳細は後述）が行われる。その結果、英語倍数表現が英文字列中に存在すれば１０進数生成手段２１が１０進数を生成する。そこで処理後の英文字列について１０進数が存在するかどうか調べ（ステップ５）、存在すれば、１０進数を図６～図８に示す日本語倍数表現処理（詳細は後述）を施し（ステップ６）、以上を終了までくり返す（ステップ７）。

【００１４】次にステップ４の倍数表現処理について図５に基づき説明する。ここで図５では倍数表現英単語Ａ$として「ｈｕｎｄｒｅｄ」を採用した処理を示しているが、他の倍数表現英単語についても同様の処理により対応できる。

【００１５】さてステップ１１にて、英文字列に「ｈｕｎｄｒｅｄ」が存在するかどうか調べる。なければ「ｈｕｎｄｒｅｄ」についての処理を終える。一方、「ｈｕｎｄｒｅｄ」が存在すれば、「ｈｕｎｄｒｅｄ」の前にＣＡＲＤ（基数詞）があるかどうか調べる（ステップ１２）。なければ「ｈｕｎｄｒｅｄ」をそのまま「ひゃく（百）」又は「１００」と訳せばよいから「ｈｕｎｄｒｅｄ」の処理を終え、あれば「ｈｕｎｄｒｅｄ」の前のＣＡＲＤを変数ｓｔａに格納する（ステップ１３）。また一旦「ｈｕｎｄｒｅｄ」を変数ｅｎｄに格納する（ステップ１４）。次にステップ１５にて、「ｈｕｎｄｒｅｄ」の次がＣＡＲＤかどうか調べる。ＣＡＲＤであれば「ｈｕｎｄｒｅｄ」の次のＣＡＲＤを変数ｅｎｄに格納し（ステップ１６）、ステップ１７へ進む。この場合、図３（ａ）のＮｏ．３のパターンとなる。一方、ステップ１５の判定が「否」であれば、「ｈｕｎｄｒｅｄ」の次が”ａｎｄ”であるかどうか調べる（ステップ１８）。”ａｎｄ”でなければステップ１７へ進む。”ａｎｄ”であれば”ａｎｄ”の次がＣＡＲＤかどうか調べる（ステップ１９）。そうであれば”ａｎｄ”の次のＣＡＲＤを変数ｅｎｄに格納する（ステップ２０）。そしてステップ１７にて、変数ｓｔａから変数ｅｎｄまでを、１０進数生成手段２１が１０進数に変換する。この変換は、”ａｎｄ”を和記号（＋）におきかえるとともに、ＣＡＲＤ×（Ａ$に対応する倍数Ｂ）という演算により行う。

【００１６】ここで図９（ｂ）のトークンリストの例では、「ｈｕｎｄｒｅｄ」が存在し（ステップ１１）、「ｈｕｎｄｒｅｄ」の前は「ｔｗｅｎｔｙ－ｔｗｏ」と

いうＣＡＲＤであるから（ステップ１２）、変数ｓｔａに「ｔｗｅｎｔｙ－ｔｗｏ」というＣＡＲＤが格納される（ステップ１３）。そして一旦「ｈｕｎｄｒｅｄ」が変数ｅｎｄに格納され（ステップ１４）、「ｈｕｎｄｒｅｄ」の次はＣＡＲＤでなく”ａｎｄ”であるから（ステップ１５，１８）、”ａｎｄ”の次の「ｆｉｖｅ」というＣＡＲＤが変数ｅｎｄに格納される（ステップ１９，２０）。そしてステップ１７の１０進数変換が行われるのであるが、この変換により、変数ｓｔａから変数ｅｎｄまでの「ｔｗｅｎｔｙ－ｔｗｏ　ｈｕｎｄｒｅｄ　ａｎｄ　ｆｉｖｅ」という表記が「２２×１００＋５」＝「２２０５」という１０進数の文字列に置き換えられる。

【００１７】次にステップ６の１０進数を日本語の倍数表現に変更する日本語倍数表現処理について、図６～図８を参照しながら説明する。図６では「億」以上の桁の処理を示す。まずステップ３０にて、日本語倍数表現生成手段２２は１０進数を入力する。次にステップ３１にて、日本語倍数表現配列ｗｈｅｒｅ〔ｉ〕（以下単に配列ｗｈｅｒｅ〔ｉ〕という）のカウンタｉに０をセットし、倍数表現付加フラグｆｌｇ（以下単にフラグｆｌｇという）に０（付加しない）をセットし、注目桁変数Ｃ（以下単に変数Ｃという）を先頭桁にし、残り文字列長変数ｌｅｎ（以下単に変数ｌｅｎという）を１０進数の全桁数とする。そして、ステップ３２にて変数ｌｅｎが８を越えているか調べ、越えていれば１億以上の数値でありステップ３３へ、越えなければステップ４０へ移る。

【００１８】ステップ３３では、フラグｆｌｇに１をセットし、配列ｗｈｅｒｅ〔ｉ〕に変数Ｃを代入し、カウンタｉをインクリメントする（ステップ３３～３５）。そして変数ｌｅｎが４の倍数であれば配列ｗｈｅｒｅ〔ｉ〕にカンマ”，”を代入する（ステップ３６，３７）。以下この処理をカンマ処理という。ここで、カンマは４桁おきに付与されるので、このように変数ｌｅｎが４の倍数かどうかチェックしている。次いでステップ３４にて１０進数の１桁分の処理がすんだので、変数ｌｅｎをデクリメントし（ステップ３８）、注目する桁を１桁くり下げ（ステップ３９）、ステップ３２に戻る。

【００１９】一方ステップ４０では、変数ｆｌｇが１かどうか調べ、１でなければ図７の処理へ移る。１であれば、億以上の桁が存在したため、ステップ３３においてフラグｆｌｇに１がセットされたものであり、配列ｗｈｅｒｅ〔ｉ〕の現在の処理中の桁に「億」をセットし（ステップ４４）、カウンタｉをインクリメントして（ステップ４２）、図７の処理へ移る。

【００２０】図７のステップ４３では、フラグｔｏｐを１（ゼロ非表示）とし、ステップ４４ではフラグｆｌｇを０（倍数表現を付加しない）としてステップ４５へ移る。ステップ４５では変数ｌｅｎが４を越えているかどうか調べる。越えているならば、残りの桁が万以上であることになる。そしてステップ４６では、フラグｔｏｐが１でありかつ変数Ｃが０であるかどうか調べる。これが是であれば、変数ｌｅｎをデクリメントし（ステップ４７）、変数Ｃを１桁下げて（ステップ４８）ステップ４５へ戻る。この処理により、「億」に直接後続する０を省略することができる。

【００２１】ステップ４６の判断が否であれば、フラグｔｏｐを０（ゼロ表示）とし（ステップ５２）、フラグｆｌｇに１をセットし（ステップ５３）、配列ｗｈｅｒｅ〔ｉ〕の現在注目している桁に変数Ｃをセットし（ステップ５４）、カウンタｉをインクリメントする（ステップ５５）。そしてステップ５６～５８において前述したカンマ処理（ステップ３６～３７と同様）を行ってステップ４７へ戻る。一方、ステップ４５において変数ｌｅｎが４を越えないときは、残りの桁が万を越えていないものであり、ステップ４９にてフラグｆｌｇが１（倍数表現を付加する）ならば、配列ｗｈｅｒｅ〔ｉ〕の現在注目している桁に「万」をセットし（ステップ５０）、カウンタｉをインクリメントして（ステップ５１）、図８の処理へ移る。一方、ステップ５９にてフラグｆｌｇが１ならばそのまま図８の処理へ移る。次に図８では千以下の桁について図７と同様の処理が行われる（ステップ５９～６９）。以上の処理が、変数ｌｅｎが０になるまで行われたら図４のステップ７へ戻る。

【００２２】なお、本実施例では、原文を英語として本発明を説明したが、英語と同様な文法形態を有する外国語であれば本発明の効果が得られることは言うまでもない。

【００２３】

【発明の効果】本発明の形態素解析装置は、原文文字列を入力する入力手段と、入力手段から入力された原文文字列を記憶する文字列記憶手段と、文字列記憶手段に記憶された原文字列をトークン毎に分割するトークン分割手段と、トークン分割手段により分割されたトークンを記憶するトークン記憶手段とを備え、原語倍数表現を記憶する倍数表現記憶手段と、トークン記憶手段に記憶されたトークンと倍数表現記憶手段に記憶された原語倍数表現とを参照して、数値を表現するトークンから、そのトークンに相当する１０進数を生成する１０進数生成手段とを有するので、原語特有の倍数表現を１０進数に変換し、正しく形態素解析することができる。

【図面の簡単な説明】

【図１】本発明の一実施例における形態素解析装置のブロック図

【図２】本発明の一実施例における形態素解析装置の機能ブロック図

【図３】（ａ）本発明の一実施例における英語倍数表現パターンの構成図

（ｂ）本発明の一実施例における倍数表現英単語とその

倍数との関係図

【図４】本発明の一実施例における形態素解析装置のフローチャート

【図５】本発明の一実施例における形態素解析装置のフローチャート

【図６】本発明の一実施例における形態素解析装置のフローチャート

【図７】本発明の一実施例における形態素解析装置のフローチャート

【図８】本発明の一実施例における形態素解析装置のフローチャート

【図９】（ａ）本発明の一実施例における形態素解析装

置の処理過程説明図

（ｂ）本発明の一実施例における形態素解析装置の処理過程説明図

（ｃ）本発明の一実施例における形態素解析装置の処理過程説明図

【符号の説明】

１０　入力手段

１３　文字列記憶手段

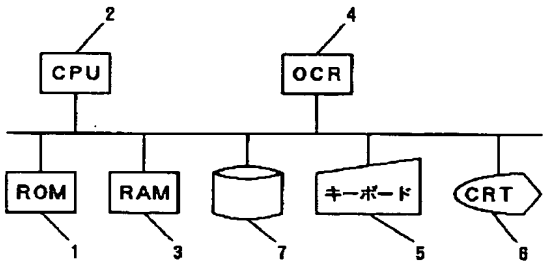
１４　トークン分割手段

１５　トークン記憶手段

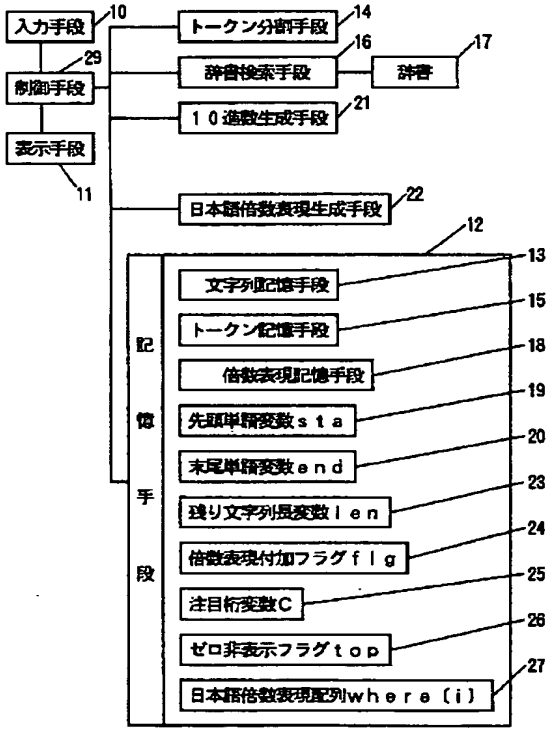
１８　倍数表現記憶手段

２１　１０進数生成手段

【図１】



【図２】



【図３】

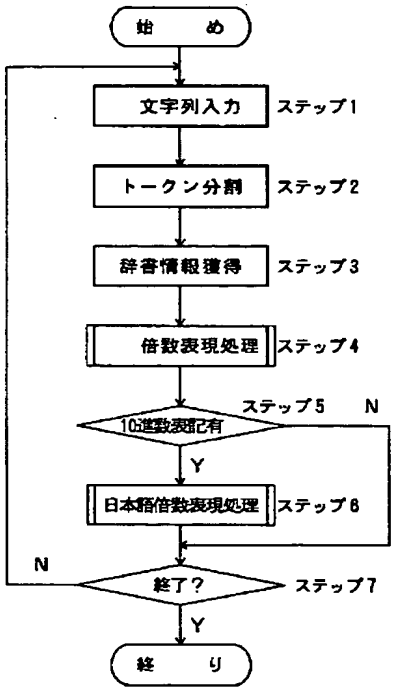
(a)

No.	英語倍数表現パターン
1	CARD　A$
2	CARD　A$+"th"
3	CARD　A$ CARD
4	CARD　A$ "and" CARD
5	CARD　A$ "and" ORD

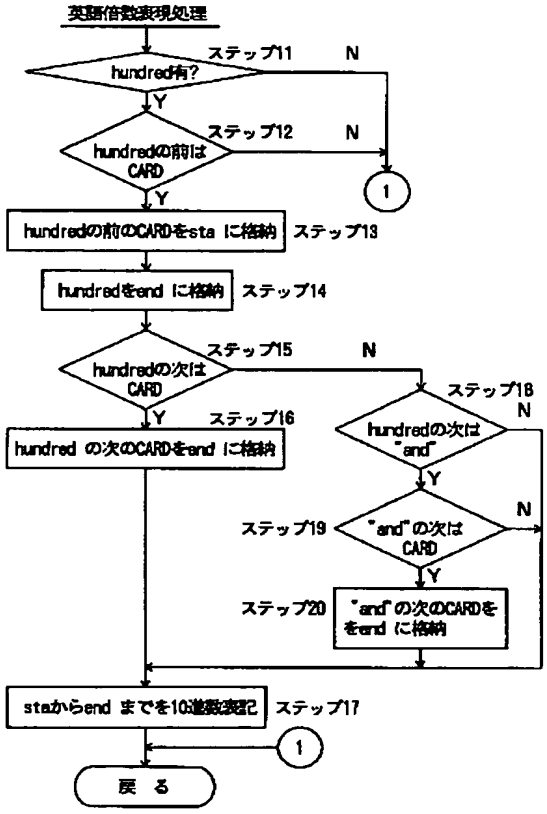
(b)

No.	倍数表現英単語Ａ$	倍　数　B
1	hundred	100
2	thousand	1000
3	million	1000000
・	・	・
・	・	・
・	・	・
N	・	・

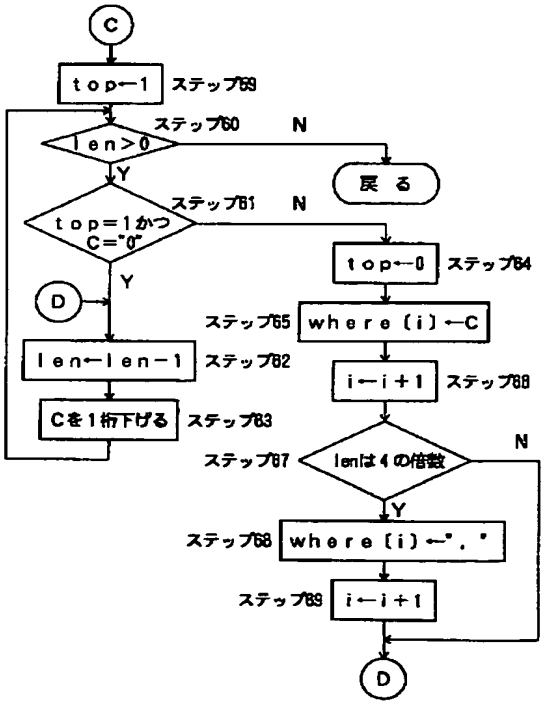
【図４】



【図５】

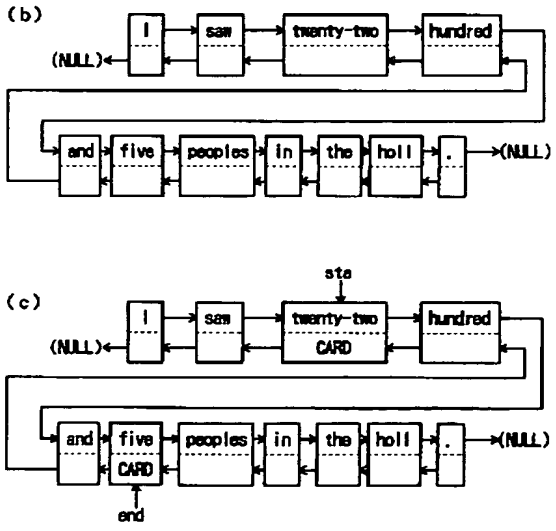


【図８】

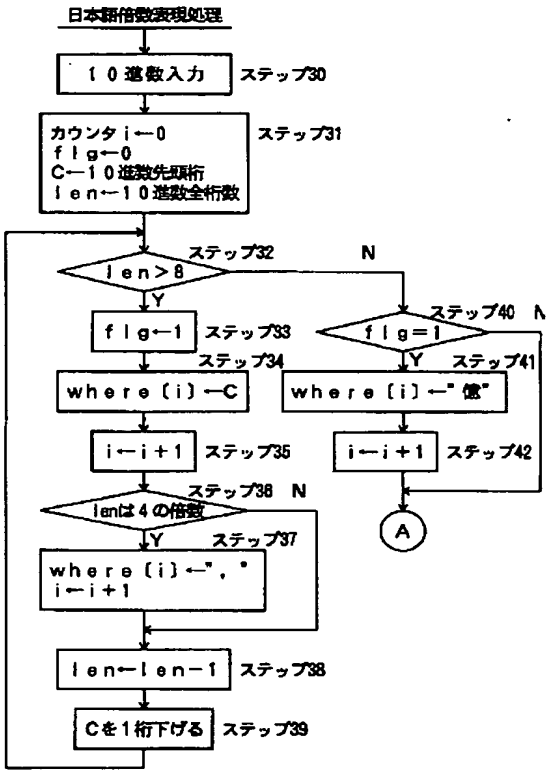


【図９】

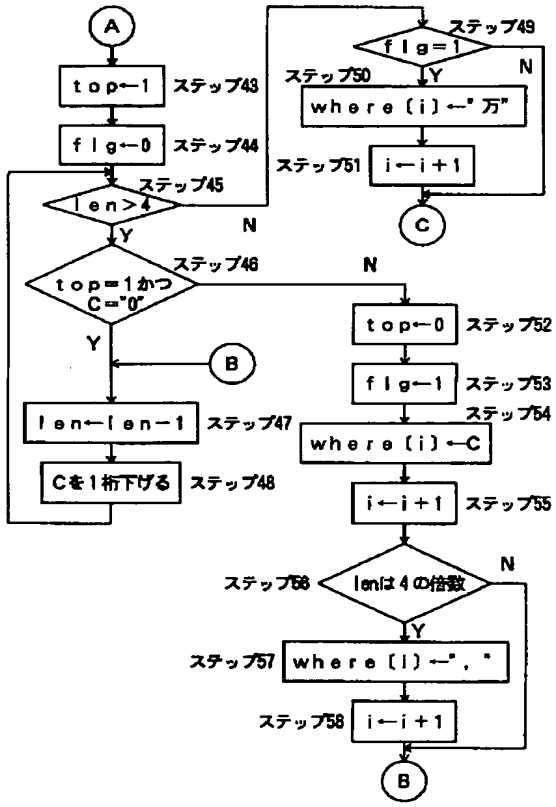
(a) I saw twenty-two hundred and five peoples in the holl.



【図６】


日本語倍数表現処理
１０進数入力　ステップ30
カウンタｉ←0
ｆｌｇ←0
Ｃ←１０進数先頭桁
ｌｅｎ←１０進数全桁数　ステップ31
ｌｅｎ＞8　ステップ32
N
Y
ｆｌｇ←1　ステップ33
ステップ34
ｗｈｅｒｅ〔ｉ〕←Ｃ
ｉ←ｉ＋1　ステップ35
lenは４の倍数　ステップ36
N
Y
ステップ37
ｗｈｅｒｅ〔ｉ〕←"，"
ｉ←ｉ＋1
ｌｅｎ←ｌｅｎ－1　ステップ38
Ｃを１桁下げる　ステップ39
ｆｌｇ＝1　ステップ40
N
Y
ステップ41
ｗｈｅｒｅ〔ｉ〕←"億"
ｉ←ｉ＋1　ステップ42
A

【図７】


A
ｔｏｐ←1　ステップ43
ｆｌｇ←0　ステップ44
ステップ45
ｌｅｎ＞4
N
Y
ステップ46
ｔｏｐ＝1かつ
Ｃ＝"0"
N
Y
B
ｌｅｎ←ｌｅｎ－1　ステップ47
Ｃを１桁下げる　ステップ48
ステップ49
ｆｌｇ＝1
N
Y
ステップ50
ｗｈｅｒｅ〔ｉ〕←"万"
ステップ51
ｉ←ｉ＋1
C
ステップ52
ｔｏｐ←0
ｆｌｇ←1　ステップ53
ステップ54
ｗｈｅｒｅ〔ｉ〕←Ｃ
ｉ←ｉ＋1　ステップ55
ステップ56
lenは４の倍数
N
Y
ステップ57
ｗｈｅｒｅ〔ｉ〕←"，"
ステップ58
ｉ←ｉ＋1
B